VERIFIABLE UNLEARNING ON GRAPH NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

VERIFIABLE PROBABILITY

Suppose the predictions of target nodes are the same as the expectation, the probability of the unlearning algorithm does not forget a real edge, named verification probability, is

$$Pr[H_1 \, is \, real] = \frac{n}{k+n}, \tag{1}$$

where $k = |E_{bkd}|$ denotes the edges added by backdoored triggers, n denotes how many edges that the request wants to forget, and H_1 is the setting in Paper "Athena: Probabilistic Verification of Machine Unlearning", section 3.1. Further, if the unlearning algorithm only forgets m real edges, the verification probability is

$$Pr[H_1 \, is \, real] = \prod_{i=0}^{m-1} \frac{n-i}{k+(n-i)}. \tag{2}$$

Intuitively, when $k >> n$, the verification probability is smaller. We thus define a ϵ-verification that should satisfies that

$$Pr[H \, is \, real] \leq \epsilon. \tag{3}$$